# 文章表現とネットワーク構造の分析に基づく大規模ＣＧＭデータ分析手法の提案

| | |
|---|---|
| 著者 | 佐藤 哲司 |
| 発行年 | 2012 |
| その他のタイトル | Analytical Methods for Huge Capacity of Consumer Generated Contents using Inner-Article Impression and Inter-Article Network |
| URL | http://hdl.handle.net/2241/118492 |

様式Ｃ－１９

# 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年 ５月 ３１ 日現在

機関番号：１２１０２
研究種目：基盤研究（Ｃ）
研究期間：2009 ～ 2011
課題番号：２１５００００９１
研究課題名（和文）
文章表現とネットワーク構造の分析に基づく大規模ＣＧＭデータ分析手法の提案
研究課題名（英文） Analytical Methods for Huge Capacity of Consumer Generated Contents using Inner-Article Impression and Inter-Article Network
研究代表者
佐藤 哲司（SATOH TETSUJI）
筑波大学・図書館情報メディア系・教授
研究者番号：70396117

研究成果の概要（和文）：ブログやツイッターなどのＵＧＣ(User Generated Contents)は，複数の著者によって断片的に書かれた記事の集合体である．本研究では，大量な記事集合における、記事間の関連性や類似性を手がかりに，記事間や著者間の経時的な変容の解明手法、文章の印象評価手法を確立するとともに，それらの知見に基づく新たな情報探索手法を考案した．質問回答サイトのアーカイブデータ等を用いた評価を行い，考案手法の有効性を確認した．

研究成果の概要（英文）：User generated contents i.e. Blogs and Twitters are written by many authors and are recently growing as a bigger part of social information sharing. In this research, we have developed several analyzing methods using both an impression of each articles and a relationship among the articles. And also we have proposed new type of navigational information retrieval methods. Our evaluation using the archive data of community question and answering site shows effectiveness of our proposed ones.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2009 年度 | 1,300,000 | 390,000 | 1,690,000 |
| 2010 年度 | 1,000,000 | 300,000 | 1,300,000 |
| 2011 年度 | 1,100,000 | 330,000 | 1,430,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,400,000 | 1,020,000 | 4,420,000 |

研究分野：総合領域
科研費の分科・細目：情報学 ・ メディア情報学・データベース
キーワード：情報検索・ウェブ情報空間・チャンク構造・文章評価・ネットワーク分析

## １．研究開始当初の背景

生活者が自ら情報を発信し，相互活用するＵＧＣ(User Generated Contents)は，元の記事にリンクを張るあるいはコメントを付記するなどの二次的な記述によって，読者も著者の一員として著作活動に参画でき，近年著しく利用が拡大している．このようなＵＧＣでは，断片的内容を記述した幾つかの記事を相互に関連させ，一塊の記事群とすることで，意味・内容が完結するチャンクを形成することができる．例えば，個人の意見や主張を発信し易くすることを目的とするブログでは，単に情報発信が容易となるだけではなく，発信された情報（話題）に対して補足や賛同，反論などを様々に引用しながら書くことができる．また，質問回答サイトでは，情報発信を質問か回答かのいずれかに限定する単純なモデルとすることで，議論の流れを把握しやすくしている．質問によって話題を提供し，その話題に対してのみ回答を積み重ねるというモデルは，質問間にまたがる議論が抑止され，回答は一人一回とするなどの制

約とあいまって議論が紛糾することもほとんどない．しかし，このモデルでは，類似した質問が繰り返し投稿される，複数の話題に関係する質問や回答を行うことが難しいなどの問題もあると考えられる．

一方，膨大に蓄積されたインターネット情報資源から，ユーザの情報要求に合致する情報を的確に検索する検索エンジンの高度化に対する要求も高まっている．この要求に応える従来研究の多くは，著者が作成したページを処理の基本単位としており，記事の断片化が進むＵＧＣを対象にすると，大量の類似記事に埋もれて検索精度が低下する，検索結果だけでなく周辺にある関連記事も読まないと内容が把握できないなどの課題も残されていた．

２．研究の目的

(1) 様々な立場の著者(情報発信者)によって付与された記事間の明示的なリンク，あるいは，テキストの類似性などの言語的な特徴によって示される暗黙的なリンクを抽出し，それらによって形成されるチャンク内およびチャンク間のネットワーク構造を分析するとともに，分析結果に基づいて効率的な情報探索を実現する手法を確立する．

(2) 記事を投稿する著者に着目し，関連するチャンクに投稿する著者間の関係性をネットワーク分析によって明らかにする．

(3) 記事中に出現する語彙の親しみ易さ（親密度）や印象を用いた文章表現の特徴抽出法を確立する．

３．研究の方法

(1) 典型的な UGC として，関連記事や続報が多い新聞記事と，質問毎にチャンクが形成される質問回答サイトの記事集合とを対比させながら，記事間にハイパーリンク構造を生成する手法を検討する．次に，得られたハイパーリンクを情報探索に適したナビゲーションネットワークへと変換する手法についても検討を深める．以上の 2 段階の技術を確立することで，UGC 記事集合を探索するナビゲーション手法を実現する．

(2) 複数の回答者が繰り返し回答する質問回答サイトのデータを用いて，回答者間ネットワークを構築する手法を考案する．一つの質問に複数の回答者が回答することと，同一の回答者が異なる質問や異なるカテゴリで回答する現象に基づいて回答者間を関連付けることで，回答者やカテゴリの特性を明らかにする．

(3) 新聞記事および質問回答サイトの複数の記事を対象に，出現する語彙の親密度と印象度を評価し，これらの分布が記事の種別やカテゴリに依存していることを明らかにする．

４．研究成果

(1) UGC 文書空間をナビゲーションによって網羅的に探索するナビゲーション手法を提案した(図 1.)．提案法では，対象とする文書集合中の記事間を関連付ける共起語をノードとするグラフを抽出した後に，ユーザによるナビゲーション(ノード間遷移)が容易となるようにノードの出次数を制限しつつ有向グラフに変換する．一般に，ノードの出次数を制限することでグラフの平均距離は長くなる傾向を示すが，平均距離の増加を抑えつつ出次数を制限できる手法である(図 2.)．

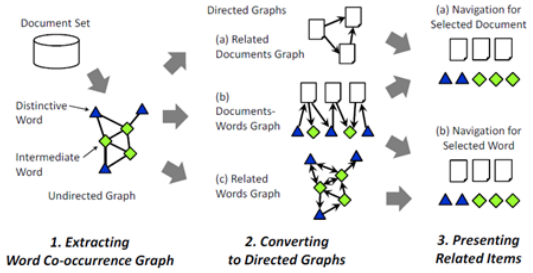


図 1. UGC 文書空間ナビゲーション手法

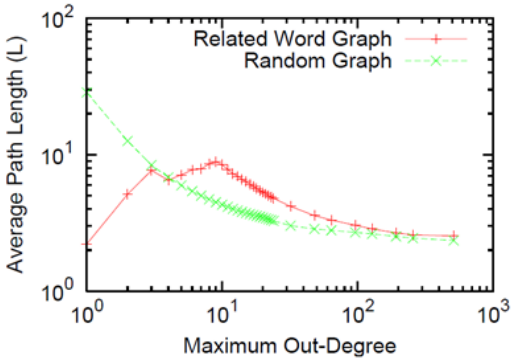


図 2. 質問回答サイト記事における平均距離

(2) 質問回答サイトにおける質問毎の平均的な回答数はカテゴリに依存する(図 3.)．また，我先にと回答する第一回答者の比率も，カテゴリに依存していることが明らかとなった(図 4.)．複数のカテゴリで回答する回答者の分布(図 5.)の結果とも合わせると，質問

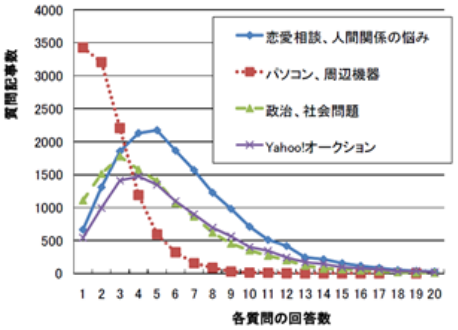


図 3. 質問毎の平均回答数

回答サイトと一言で言っても，カテゴリ毎に異なる性質を備えたコミュニティが形成さ

れていることを強く示唆している．

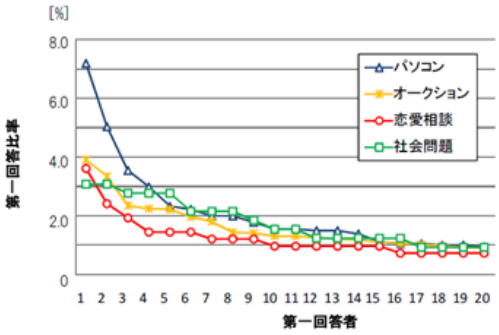


図 4. 第一回答者のカテゴリ比較

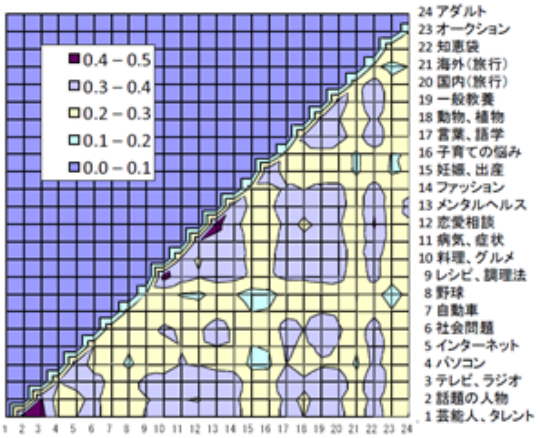


図 5. カテゴリ間に拡がる回答者の興味

「優れた回答者が回答した質問でベストアンサーを取得した回答者はより優れている」とする考えに基づいて，個々の回答者の役割と貢献度を定量的に算出する QARank，QAHits アルゴリズムを提案した．ベスト回答者に対する QARank 値の分布(図 6.)から，PC カテゴリには 10 名程度の優れた回答者が存在すること，恋愛カテゴリには特定の回答者にベスト回答が偏らないことが明らかとなった．同様の評価を第一回答者についても実施し，カテゴリ毎に特徴が異なることを確認した．

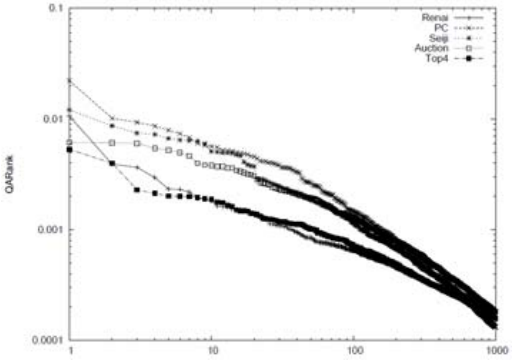


図 6. ベスト回答者の QARank 値分布

(3) 新聞記事および質問回答サイトの複数の記事を対象に，出現する語彙の親密度と印象度を評価し，これらの分布が記事の種別やカテゴリに依存していることを明らかにした．語彙特性辞書から文章の印象評価に有効と考えられる語を抽出し，利用者実験によって 9 種類の因子に分類した(表 1.)．

この結果を用いて文章評価を行ない，質問と回答の文章中に含まれる印象語が，ベストアンサーの推定に一定の効果があるとの示唆を得た．更に，ある単語から喚起されるイメージがどの程度思い浮かべやすいかを表す単語心像性を特徴量に加えることで推定精度を向上できることを示した．

表 1. 因子と対応する印象語

| 因子 | 印象語 | | |
|---|---|---|---|
| 第1因子（的確性） | 説得力がある<br>素晴らしい<br>真実味がある<br>充実した<br>丁寧な | 流暢な<br>好ましい<br>清々しい<br>美しい | 重要な<br>巧みな<br>妥当な<br>的確な |
| 第2因子（不快性） | 不快な<br>残念な<br>幻滅した | 憤慨した<br>不当な<br>怖い | 非常識な<br>呆れる |
| 第3因子（独創性） | 独創的な<br>斬新な | 予想外な<br>不思議な | 特殊な |
| 第4因子（容易性） | 易しい | 明瞭な | 難しい |
| 第5因子（執拗性） | 細かい | しつこい | 長い |
| 第6因子（曖昧性） | 曖昧な | 不充分な | |
| 第7因子（感動性） | 心温まる | 感動的な | |
| 第8因子（努力性） | 涙ぐましい | | |
| 第9因子（熱烈性） | 熱い | 力強い | |

(4) 質問回答サイトに投稿された質問記事は，自然言語で記述されたユーザの情報要求であることから，この質問記事を用いてインターネット情報検索における多様なクエリ拡張を実現する方法を考案した．クエリ拡張は，質問を投稿するカテゴリと投稿時期の 2 次元でファセットを構成したタグクラウドで行い，ユーザが選択したキーワードに基づいて拡張クエリを生成すると共に，拡張の根拠となる質問記事を提示する(図 7.)．タグクラウドに提示する語は，LDA 手法を用いて各ファセットにおける重要語を抽出している．


図 7. タブ切替による 2 次元ファセット

一方の季節性は，季節(四季)毎の変動係数(C.V)を算出し，バーストを判定することで

特徴的な語を抽出する手法を考案した．質問回答サイトに投稿された 3 年間の質問を対象に分析を行ない(図 8.)，季節性の高い語を抽出できることを検証した．

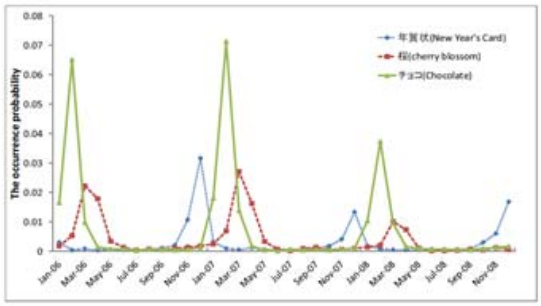


図 8. 季節性の高い語の出現確率の推移

(5) ユーザの情報要求と情報資源(ウェブページ)との間で情報粒度が異なることは，情報探索において大きな課題となっている．情報要求の粒度が大きい場合への対処法として，検索結果となる複数ページ間の関係性を分析するシステム(図 9.)を考案した．検索結果を特徴付ける語を LDA 等の手法を用いて抽出し，その語を軸キーワードとすることで，検索結果を 2 次元空間に配置する．インタラクティブに軸キーワードを入れ替えながら所望の情報を得る利用者実験を行ない，多様な検索結果を網羅する必要がある情報探索は有効であることを確認した．


図 9. ページ間の関係性に基づく検索手法

一方，ユーザが要求する情報がページ内の限定領域にのみ存在する場合には，ページ内の領域を切り出して，集約・編集できる手法が有効となる．そこで，ページを階層的な構造とテキスト量に基づいて複数のブロックに分割する手法を考案した．ユーザが指定したブロックを単位としてスクラップブックを作成できるシステム(図 10.)を実装し，利用者実験によって有効性を確認した．

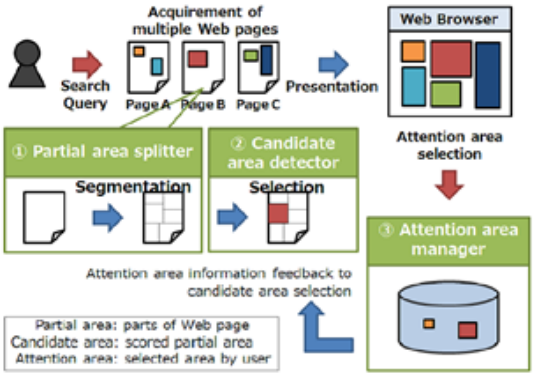


図 10. ブロックを単位とする情報集約

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 10 件）


① Y. Yokoyama, T. Hochin, H. Nomiya, T. Satoh: Improvement of Estimation Accuracy of Factor Scores from Feature Values of Statements, IIAI/ACIS International Symposium on Innovative E-Services and Information Systems (IEIS 2012) (May, 2012) 査読有り, DOI: 10.1109/ICIS.2012.60

② A. Otsuka, Y. Seki, N. Kando, Tetsuji. Satoh: QAque: Faceted Query Expansion Techniques for Exploratory Search using Community QA Resources, Proc. of the 21th International Conference companion on World Wide Web, WWW'12 Companion (CQA'12), pp. 799-806 (Apr. 2012), 査読有り, DOI: 10.1145/2187980.2188203

③ Y. Tasaki, T. Fukuhara, T. Satoh: Aggnel: An information aggregation system of partial contents from multiple Web pages, Proc. of the 2012 26th International Conference on Advanced Information Networking and Applications Workshops, AINA Workshop 2012, pp. 815-820 (Mar. 2012), 査読有り, DOI:10.1109/WAINA.2012.213

④ Y. Yokoyama, T. Hochin, H. Nomiya, T. Satoh: Obtaining Factors Describing Impression of Questions and Answers and Estimation of their Scores from Feature Values of Statements, Proc. of 1st ACIS International Symposium on Software and Network Engineering (SSNE2011) (Dec. 2011), 査読有り, DOI:10.1109/ICIS.2012.60

⑤ 大塚 淳史, 関 洋平, 神門 典子, 佐藤 哲司: 情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討, 情報処理学会 論文誌 データベース (TOD) , Vol.4, No.3, pp. 1 -11 (Oct. 2011), 査読有り, http://db-event.jpn.org/deim2011/pr

oceedings/pdf/f6-3.pdf

⑥ 島田諭，福原知宏，佐藤哲司：出次数制約付き有向グラフを用いた関連語による文書空間ナビゲーション手法，情報処理学会 論文誌，Vol. 52，No. 4，pp. 1831-1842（Apr. 2011），査読有り，http://ci.nii.ac.jp/naid/110008508013

⑦ 横山友也，宝珍輝尚，野宮浩揮，佐藤哲司：質問回答サイトの質問文と回答文の印象評価とベストアンサーの推定，日本感性工学会 論文誌，Vol. 10，No. 2，pp. 221-230（Mar. 2011），査読有り，DOI:10.5057/jjske.10.221

⑧ 島田 諭，福原知宏，佐藤哲司：文書空間ナビゲーションのための出次数制約付き有向グラフ生成手法，情報処理学会 論文誌 データベース（TOD），Vol.3，No.2，pp. 111-122（July 2010），査読有り，http://ci.nii.ac.jp/naid/110007990086

⑨ 佐藤弘樹，島田諭，伏見卓恭，福原知宏，斉藤和巳，佐藤哲司：質問回答サイトにおける QARank を用いたユーザ貢献度の推定法，情報社会学会誌，Vol.4，No.2，pp. 5-13（Mar.2010），査読有り，http://infosocio.org/vol4no2-1.pdf

⑩ S. Shimada，T. Fukuhara，T. Satoh：S-node： A Small-World Navigation System for Exploratory Search，Proc. of the 5th Asia Information Retrieval Symposium，AIRS2009，LNCS 5839，Information Retrieval Technology，pp. 420 - 431（Oct. 2009），査読有り，DOI：10.1007/978-3-642-04769-5_37

〔学会発表〕（計 44 件）

① 枝 隼也，島田 諭，関 洋平，神門 典子，佐藤 哲司：複数人での Web 協調探索のための探索履歴可視化手法の提案，電子情報通信学会 他共催，第 4 回データ工学と情報マネジメントに関するフォーラム DEIM2012 論文集，C9-1（Mar. 5，2012）.

② 大塚 淳史，関 洋平，神門 典子，佐藤哲司：コンテキスト切替による多様な情報要求に対する Web 検索手法の提案，電子情報通信学会 他共催，第 4 回データ工学と情報マネジメントに関するフォーラム DEIM2012 論文集，F8-4（Mar. 4，2012）.

③ 島田 諭，山口 裕太郎，佐藤 哲司：マイクロブログにおける情報伝搬距離に着目したユーザプロファイリング，電子情報通信学会 他共催，第 4 回データ工学と情報マネジメントに関するフォーラム DEIM2012 論文集，D8-5（Mar. 4，2012）.

④ 杉本 和香奈，佐藤 哲司：既存レシピを活用した並行調理スケジュール法の提案と評価，電子情報通信学会 他共催，第4回データ工学と情報マネジメントに関するフォーラム DEIM2012 論文集，E8-1（Mar. 4，2012）.

⑤ 林 大策，福原 知宏，佐藤 哲司：情報整理を支援する対話型検索インタフェースの提案と評価，電子情報通信学会 他共催，第 4 回データ工学と情報マネジメントに関するフォーラム DEIM2012 論文集，E7-1（Mar. 4，2012）.

⑥ 香川 雄一，島田 諭，神門 典子，佐藤哲司：コミュニティ QA における質問・回答が互いに及ぼす影響の可視化に関する検討，電子情報通信学会 他共催，第4回データ工学と情報マネジメントに関するフォーラム DEIM2012 論文集，C3-3（Mar. 3，2012）.

⑦ 山口 裕太郎，島田 諭，佐藤 哲司：人物の呼称を用いたマイクロブログ記事検索に関する一検討，電子情報通信学会他共催，第 4 回データ工学と情報マネジメントに関するフォーラム DEIM2012 論文集，F2-4（Mar. 3，2012）.

⑧ 山本 修平，佐藤 哲司：Twitterからの実生活情報の抽出法の提案，電子情報通信学会 他共催，第 4 回データ工学と情報マネジメントに関するフォーラム DEIM2012 論文集，F2-3（Mar. 3，2012）.

⑨ 横山 友也，宝珍 輝尚，野宮 浩揮，佐藤哲司：文末表現を考慮した文章の特徴量を用いた質問回答文の因子得点の推定，第 17 回公開シンポジウム「人文科学とデータベース」予稿集，pp. 9-20（2012.1.7）.

⑩ 横山 友也，宝珍 輝尚，野宮 浩揮，佐藤哲司：単語心像性を用いた質問回答文の因子得点の推定精度の向上，平成 23 年度情報処理学会関西支部支部大会，C-13（2011.9.22）.

⑪ 田崎雄一郎，福原知宏，佐藤哲司：複数 Web ページの注目領域を対象とした情報探索と集約手法の提案，情報アクセスシンポジウム 2011(IAS2011)，pp. 17-22（Sept. 7，2011），査読有り.

⑫ 枝 隼也，福原 知宏，佐藤 哲司：話題範囲に着目した Web 閲覧履歴の空間的把握手法の提案，情報処理学会，マルチメディア，分散，協調とモバイルシンポジウム DICOMO2011，pp. 1507-1512（July 8，2011），査読有り.

⑬ 大塚淳史，関洋平，神門典子，佐藤哲司，情報要求の言語化支援のためのコンテキスト提示型クエリ拡張法の提案と評価，情報処理学会，マルチメディア，分散，協調とモバイルシンポジウム

DICOMO2011, pp. 22-29 (July 5, 2011), 査読有り.
⑭ 横山友也, 宝珍輝尚, 野宮浩揮, 佐藤哲司 : 文章の特徴量を用いた質問回答文の因子得点の推定精度の向上, 日本感性工学会関西支部大会, 4p. (May 20-21, 2011).
⑮ 渡邊直人, 島田諭, 関洋平, 神門典子, 佐藤哲司: コミュニティ QA における質問の多面的評価法の検討, 情報知識学会第 19 回（2011 年度）年次大会, 情報知識学会誌, Vol. 21, No. 2, pp. 163-168 (May 28, 2011).
⑯ 林 大策, 福原知宏, 佐藤哲司: 軸キーワードによる観点選択を実現したインタラクティブ検索の提案, 電子情報通信学会, 第 19 回 Web インテリジェンスとインタラクション研究会, WI2-2011-18, pp. 79-80 (Mar. 8, 2011).
⑰ 田崎雄一郎, 島田 諭, 福原知宏, 佐藤哲司: Web ページからの注目領域抽出に基づく横断型情報閲覧システム, 電子情報通信学会, 第 19 回 Web インテリジェンスとインタラクション研究会, WI2-2011-20, pp.83-84 (Mar. 8, 2011).
⑱ 枝 隼也, 福原知宏, 佐藤哲司: Web 閲覧履歴の空間的把握手法の提案, 電子情報通信学会, 第 19 回 Web インテリジェンスとインタラクション研究会, WI2-2011-05, pp. 27-28 (Mar. 7, 2011).
⑲ 横山 友也, 宝珍 輝尚, 野宮 浩揮, 佐藤哲司: 文章の特徴量を用いた質問回答文の因子得点の推定, 日本感性工学会, 第 6 回日本感性工学会春季大会 22D-02, 4p (Mar. 4, 2011).
⑳ 新谷歩生, 関洋平, 佐藤哲司: 投稿間隔に基づくマイクロブログからの話題チャンク抽出に関する一検討, 電子情報通信学会データ工学研究専門委員会 他共催, 第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM2011), A1-2 (Feb. 27, 2011).
㉑ 島田 諭, 小出 明弘, 斉藤 和巳, 佐藤哲司: QA2 部グラフにおけるモチーフを用いたコミュニティの経時的変化に関する分析, 情報社会学会, 第 3 回知識共有コミュニティワークショップ論文集, pp. 2-10 (Dec. 15, 2010), 査読有り.
㉒ 小出 明弘, 斉藤 和巳, 佐藤 哲司: モチーフによる QA 二部グラフの構造分析, 情報処理学会 共催, Web とデータベースに関するフォーラム(WebDB Forum) 2010, 4B-3 (Nov. 12, 2010), 査読有り.
㉓ 横山友也, 宝珍輝尚, 野宮浩揮, 佐藤哲司: 質問回答サイトの質問文と回答文の印象評価とベストアンサーの推定, 日本感性工学会, 第 12 回日本感性工学会大会, 2C1-4 (Sept. 12, 2010), 査読有り.
㉔ 枝 隼也, 福原 知宏, 佐藤 哲司: ユーザの Web 探索履歴からのキーワード遷移グラフの抽出法に関する一検討, マルチメディア, 分散, 協調とモバイルシンポジウム DICOMO2010 論文集, 2D-2, pp. 406 - 413 (July 7, 2010), 査読有り.
㉕ 林 大策, 佐藤 哲司: キーワード平面を用いたインタラクティブ検索の提案と評価, 情報処理学会, マルチメディア, 分散, 協調とモバイルシンポジウム DICOMO2010 論文集, 1G2, pp. 229 - 237 (July 7, 2010), 査読有り.
㉖ 岡田 仁之, 島田 諭, 福原 知宏, 佐藤哲司: Wikipedia を利用した日本語作文支援システムの開発, 情報処理学会, 人文科学とコンピュータシンポジウム(じんもんこん 2009)論文集, pp. 225 - 230 (Dec. 19, 2009), 査読有り.
㉗ 佐藤 弘樹, 島田 諭, 伏見 卓恭, 斉藤 和巳, 佐藤 哲司: 質問回答サイトにおける QARank を用いたユーザ貢献度の推定, 情報社会学会, 第 2 回知識共有コミュニティワークショップ論文集, pp. 43 - 51 (Dec. 13, 2009), 査読有り, 優秀賞.
㉘ 佐藤 弘樹, 島田 諭, 福原 知宏, 斉藤 和巳, 佐藤 哲司: 質問回答サイトにおける投稿種別に注目したコミュニティ分析手法, 情報処理学会 共催, Web とデータベースに関するフォーラム (WebDB Forum) 2009, 2A-3 (Nov. 19-20, 2009), 査読有り.
㉙ 島田諭, 福原知宏, 佐藤哲司: 文書集合の特性を考慮した包括的 Web ナビゲーション, 情報処理学会, マルチメディア, 分散, 協調とモバイルシンポジウム DICOMO2009, 1B-3, pp. 47 - 54 (July 8, 2009), 査読有り, 優秀論文賞.

６．研究組織

(1)研究代表者

佐藤 哲司（SATOH TETSUJI）
筑波大学・図書館情報メディア系・教授
研究者番号：70396117

(2)研究分担者

福原 知宏（FUKUHARA TOMOHIRO）
独立行政法人産業技術総合研究所
サービス工学研究センター・特別研究員
研究者番号：50436581
宝珍 輝尚（HOCHIN TERUHISA）
京都工芸繊維大学・工芸科学研究科・教授
研究者番号：00251984
斉藤 和巳（SAITO KAZUMI）
静岡県立大学・経営情報学部・教授
研究者番号：80379544